

STATISTICS FOR ENVIRONMENTAL MONITORING

Jennifer Brown

*Biomathematics Research Centre
Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand
Email: j.brown@math.canterbury.ac.nz*

Report Number: UCDMS2002/5

February 2002

Keywords: survey design, long-term monitoring, population abundance, benthic surveys

Introduction to Environmental Monitoring

1.1 Observational and Manipulative Experiments

Environmental studies can be broadly categorised as either:

- i) Observational experiments, or
- ii) Manipulative experiments.

An observational experiment is one where the data is collected by observing some existing process (Manly 1992). For example, the population of Hector's dolphins around Banks Peninsula was surveyed each year from 1990 to 1995 to estimate the trend in the population size once the Banks Peninsula Marine Mammal Sanctuary was put in place. Generally, observational studies are conducted where the environmental process is of interest but the mechanism driving the process may not be well understood. In contrast, a manipulative experiment is conducted when specific variables are altered and the outcome, or effect, is estimated (Manly 1992). If the monitoring of Hector's dolphins had begun well before the establishment of the Sanctuary the survey could have been considered an "experiment" in that the population trend prior to and after the Sanctuary establishment could have been compared. In this example, the protection status of the dolphins would have been the variable that was changed.

While there is a lot more control over confounding effects in designed experiments often they can not be conducted. For example, it is not possible to design an experiment to measure the effect of an oil spill in a harbour because it would be difficult to get permission to create an artificial spill.

As another example of an observational study compared with an experiment consider a study on the effects of changes in weather on the likelihood that a possum will be caught in a trap. The study could be conducted as an observational experiment. By repeatedly trapping a population of possums over time differences in the number of traps catching a possum could be related to differences in weather, such as rain or fine night, air temperature, wind direction. This study would need to be carried out over a long time period to be able to capture as much variation in weather as possible. There is a potential problem with such a design though. Any observed differences in possum-catch are likely to be confounded by changes in possum behaviour - a possum caught once maybe weary of being caught again. For example, perhaps the weather progressively deteriorated throughout the study so that at the start of the study there were fine-weather nights and bad-weather nights near the end. An observed reduction in the proportion of traps catching a possum may appear to be due to differences in weather but in fact due to possums having a learnt aversion to traps. There may be confounding factors that were not even considered. Possum-catch may reduce over time simply because the possum diet had shifted between the start and end of the study.

A manipulative experiment would involve using captured possums in pens and artificially creating different temperatures, rainfall, and wind within each pen. The study could be then carried out over one night. With such a design there would be strong evidence that the observed differences in trap-catch was an effect of different temperature, or rainfall, or wind etc. rather than some behavioural aversion.

1.2 Internal and External Validity

A criticism of the manipulative experiment using caged possums to study differences in possum-catch is that the results are valid only for caged animals. In the wild possums may display totally different behaviour. For example, possum catch may be observed to be highest with high humidity with caged animals but with natural animals it may be lowest. Such an experiment is said to have low external validity (Manly 1992). External validity is when the results could be applied to the wider population. Internal validity is when the observed effect is due to the factor being studied rather than some other (confounding) effects. The observational study of

possums in the wild will have low internal validity. Clearly a balance between the two must be met and this depends on the study objective.

1.3 Experiments and Quasi-Experiments

Another concept is the distinction between a "true" experiment and, probably what is more common in environmental science, a quasi-experiment. The idea of a true experiment developed from agricultural science where there were very controlled situations. There are three key features of an experiment:

- i) Experimental units are randomly allocated to treatments,
- ii) There is a control, and
- iii) There is replication.

For example, to study the effect of fertiliser on tree seedling growth, a coin is tossed to decide which of two nursery beds (experimental units) are to be fertilised. The bed that is not fertilised is considered the control. The control is used to find out what growth is in the absence of the treatment. Many pairs of nursery beds are used to provide replication because there are always inherent variation among experimental units.

It is rare to be able to have such control over an environmental study. A quasi-experiment is where the study has some, but not all, of the features of a true experiment (Manly 1992). The point is that an attempt should be made to have as much as possible some of the elements of a true experiment.

1.4 Pseudoreplication

Pseudoreplication is defined as:

"the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated or replicates are not statistically independent." (Hurlbert 1984).

This concept is best explained by some examples. To assess diversity in grassland following burning a 1ha study area was randomly located within a burnt and unburnt field. Within each of the 1ha areas 15 - 1m² quadrats were randomly located. Does this design give 15 independent replicate samples from the burnt and unburnt area? The answer is no if the question is about the general effect of burning. The problem is that the "replicate" quadrats are not replicated at the correct scale. The idea of replication is to provide a measure of the intrinsic variability of the area that has nothing to do with the treatment (in this example, the treatment is burning) (Underwood 1997). The observed differences in the burnt and unburnt quadrats could be due site differences and not the effect of burning. The observed differences are confounded by other differences. Underwood, in fact, uses the term confounding in preference to the term pseudoreplication because it draws attention to what is needed;

"It is not replication as such that is the problem. The difficulty is to separate out the differences among treatments that are due to the experimental factor from any differences due to other factors. The logic of the experiment requires this, so that any differences found can be unambiguously attributed to the process proposed in the model. Other differences confound such conclusions, making logical interpretation impossible." (Underwood 1997).

As another example, in a study of the diurnal pattern of Hector's dolphin in Akaroa Harbour an observational design was used. Each morning and afternoon the direction the dolphins were swimming in when they were at the harbour entrance was observed. The basic sampling unit was the pods of dolphins, not the individual dolphins. The use of the individual dolphins as the

data points would be pseudoreplication because all dolphins within a pod will be swimming in the same direction. The dolphin pods were the level at which the sampling units are independent. Using the individual dolphins as the sample unit artificially inflates the sample size.

1.5 Two levels of pseudoreplication

Extension of the statistical (inductive) conclusions from an observational design beyond the specific study areas/populations to other unstudied areas/populations is one of the common forms of pseudoreplication (Hurlbert 1984, Stewart-Oaten et al. 1986). Consider an example of the accidental spill of oil into Lyttelton Harbour. Deductive inferences concerning general conclusions of cause-and-effects of the oil (that extend beyond the specific study areas/populations) may be possible if enough independent studies of different discharges of the oil are observed to produce similar effects. However, statistical (inductive) inferences beyond the study areas/populations are not possible using a simple observational study within the one harbour.

Results of observational studies such as evaluation of environmental impacts of accidental spills of oil or chemicals are often referred to a pseudoreplication in the biological literature. This use of the word is misleading unless it is qualified by adding the comment that the subsampling of treatment and non-treatment areas/populations is pseudoreplication if statistical conclusions are extrapolated beyond the treatment and non-treatment areas/populations. In other words, random sampling within an observational study is not pseudoreplication if the statistical inferences are limited to the specific areas or populations studies. It is the actual application of inferential statistics to unreplicated treatments or dependent replicates that causes "pseudoreplication." Single replicates per treatment or dependent data are not necessarily bad, or avoidable in field studies. However, it is dangerous to extrapolate inductive conclusions from such data using inferential statistics.

A second level of pseudoreplication occurs if dependent data from the basic sampling units of observational studies are analysed as if they are independent. Essentially the "sample size" is artificially inflated by analysing more than one datum per basic sampling unit. The importance of identifying and maintaining the integrity of data from the basic sampling units cannot be overemphasised. Things can get complicated. A good rule to follow is that statistical inferences should be based on only one value from each sample unit (unless the dependent data are properly handled in the analysis). For example, if 5 quadrats are randomly located in a study area, then design/data-based statistical inferences to the area should be based on 5 values; regardless of the number of plants, organisms, split-samples, etc, which may be present and measured or counted.

As another example, heights of individual plants were recorded for all plants in randomly located quadrats within a study area. The variation from plant to plant within a quadrat is an inappropriate measure of variation for statistical comparisons of a pair of treatment and non-treatment sites. A researcher would be guilty of pseudoreplication if the within quadrat variance is used in the statistical tests to compare mean height of plants in a particular pair of assessment and control sites.

1.6 Identifying Pseudoreplication

Problems associated with incorrect identification of data from the sampling units can give rise to incorrect statistical precision of estimated end-points. A simple example of pseudoreplication occurs if a single collection of material (sediment, plant tissue, etc.) might be taken at one point in an area, and then split several times in the laboratory. Analyses of each subset of the material might be conducted. Variation among such "replicates" is proper for study of the accuracy and precision of the laboratory measurement procedures, but does not represent spatial or temporal variation in the area and/or variation due to the field sampling. Variation

among such replicates is not the correct measure of variation for comparison of assessment and control areas by statistical procedures. This example of pseudoreplication is presented because it is relatively easy to understand and has occurred in practice. The problem is usually easy to fix. In this example, simply average the results of the toxicity analyses on the subsets (split samples) to produce one number for each location of collection in the field. In general, use one datum per sampling unit from the field in the statistical procedures.

Although Hurlbert (1984) states that pseudoreplication widely occurs in both observational studies and manipulative experiments, he focuses the majority of his review on manipulative experiments. He also defines temporal pseudoreplication to occur when multiple samples are taken sequentially over time on the same sampling units (i.e., time series data), but the data are analysed as if they are independent. The experimenter should assume that the potential for false treatment effects is high because successive samples from a single unit taken over time are likely correlated. For example, in the study of growth of a weed, 30 successive measurements of the size of one patch is not equivalent to relocations of 30 randomly sampled patches from the population. Hurlbert did not discuss the use of "repeated measurement experimental designs" (the analysis of repeated measurements on the same experimental units over time). This theory has potential for solving many of the temporal pseudoreplication problems in manipulative experiments (e.g., Milliken and Johnson 1984).

Example: Sampling Vegetation Cover

Data from three transects on the proportion of bare substrate has been collected. The data were recorded in 5m sections, that is the proportion of bare substrate in the first 5m section of transect is recorded, then the proportion in the next 5m section, and so on. Two of the transects are 35m in length and the third is 50m in length. In total there are 24 - 5m sections. Some of the data is displayed in Table 1.1.

Table 1.1 Proportion of bare substrate in three transects recorded in 5m sections.

Section	Proportion bare substrate 1999			
	Transect 1	Transect 2	Transect 3	Overall
0	0.3		0.1	
5	0.5	0.1	0.85	
10	0.35	0.5	0.5	
15	0.35	0.4	0.25	
20	0.55	0.2	0.3	
25	0.6	0.05	0.8	
30	0.4	0.1	0.15	
35		0.02	0.3	
40			0.2	
45			0.3	
Transect average	0.435714	0.195714	0.375	0.335476
Transect standard deviation	0.114434	0.184739	0.260608	0.222674
Transect cv	0.262636	0.943923	0.694955	0.633838

To answer the question "Is there a change in the proportion of bare substrate over time", the appropriate level of analysis is the transect averages or totals (Figure 1.1). The proportion of bare substrate from each 5m section of transect is averaged over the number of sections within each transect. The sample size is therefore three since there are three transects. There is no obvious trend in increasing, or decreasing amounts of bare substrate over time. One test to see if the amounts change significantly among years would be to conduct a repeated measures analysis.

This analysis used the transect averages and ignores information in the variation of the bare substrate within transects. For example, consider two transects which both have on average 0.3 proportion of the surface area as bare substrate. On one transect the first seven 5m sections have no bare substrate and then the last three sections are all bare. On the other transect all ten sections have 0.3 proportion bare substrate. These two transects have quite different coverage and the management implications differ. In the first transect perhaps there has been some disturbance in the end sections of the transect, while in the second transect there may be the same level of disturbance over all the transect. To compare within-transect variation the standard deviation and cv of the proportion of bare substrate *within* each transect is calculated. The average cv over the three transects is also shown in Figure 1.1.

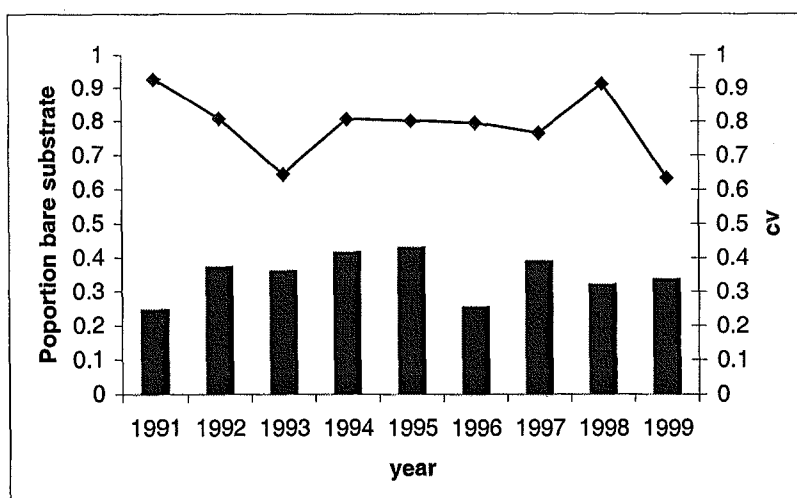


Figure 1.1 Average amount of the proportion of bare substrate in three transects from 1991 and 1999. Also shown is the average amount of within - transect variation in the proportion of bare substrate. This is measured by the cv of the 5m sections within each transect.

References

- Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54(2):187-211.
- Manly, B.F.J. (1992). *The Design and Analysis of Research Studies*. Cambridge University Press, Cambridge.
- Miliken, G.A. and Johnson, D.E. (1984). *Analysis Of Messy Data*, Volume 1: designed experiments. Lifetime Learning Publications, Blemont, California.
- Stewart-Oaten, A., W.W. Murdoch, and K.R. Parker. (1986). Environmental impact assessment: "Pseudoreplication" in time? *Ecology* 67:929-940.
- Underwood, A.J. (1997). *Experiments in Ecology*. Cambridge University Press, Cambridge.

Survey Designs for Environmental Monitoring

2.1 Introduction

The purpose of many monitoring schemes is to detect the effect of a human impact on natural populations (Underwood 1992). The appropriate analysis therefore is to measure the change in an associated environmental variable. However, natural populations display considerable temporal and spatial variation so the sampling design and analysis must be capable of distinguishing between what is normal variation and variation that may be attributed to the human impact (Skalski and McKenzie 1982, Underwood 1994).

2.2 Monitoring Objectives

The monitoring scheme chosen should relate directly to management and for this objectives must be clearly stated. Monitoring has been defined as:

"...the collection and analysis of repeated observations or measurements to evaluate changes in condition and progress towards meeting a management objective." (Elzinga *et al.* 1998)

The objectives should be stated in terms what is being measured, over what time, in what location, what measure is to be used, and what change is to be detected in the measure. For example, a management objective may be to reduce a pest species to a threshold level as measured by some index, within a certain area, within a certain time (Ringold *et al.* 1996, Gibbs *et al.* 1999). Clearly stated objectives should make the choice of design, including what sites to select for the surveys and choice of analysis techniques of the data, obvious.

As another example, in designing a monitoring scheme for Karner blue butterflies in Wisconsin, USA, the two objectives were to assess: i) the overall effectiveness of the conservation efforts on the butterfly density within Wisconsin by assessing regional trends; and ii) the effectiveness of individual conservation strategies within tracts of land to allow comparison among strategies (Brown and Boyce 1996). The timeframe for these objectives were also stated.

The first objective required data collected over time from a sample of sites that were representative of the Wisconsin State. The purpose of statistical analysis was to separate natural variation from variation and trends resulting from the conservation efforts. The second objective required data to be collected from the areas managed under the conservation strategy of interest and for this to be compared with data from control sites that are governed by natural ecological processes (Eberhardt 1976, Green 1979 p.29).

In the Galápagos Island monitoring is undertaken to assess the general goal of preservation of biological diversity. The monitoring has also been established to evaluate changes in wildlife resources. However, there are different types of changes. Natural changes are simply observed, whereas anthropogenic changes require restoration (for past changes), mitigation (for current changes) or prevention (for potential changes) (Gibbs *et al.* 1999). The monitoring has to be designed to allow these different changes to be identified.

The special requirements of environmental monitoring schemes has led to interest recently in more complicated designs that include aspects of random sampling, good spatial cover, and the gradual replacement of sampling sites over time (Skalski, 1990; Stevens and Olsen, 1991; Overton *et al.*, 1991, Urquhart *et al.*, 1993). Monitoring designs that are optimum in some sense have also attracted interest in recent years (Fedorov and Mueller, 1989; Caselton *et al.*, 1992).

2.3 Temporal Replication of Sites - Monitoring over Time

Environmental monitoring typically requires a number of years of sampling to be able to detect real biological trend (Barker and Sauer 1992). There are four general sources of pattern in population data:

- i) Trend resulting from a population change, i.e., the population trend that we are wanting to detect in monitoring;
- ii) Irregular environmental perturbations e.g., unusual weather events;
- iii) Autocorrelation due to population processes, i.e., the population size in one year is expected to be related to the population size in the previous year; and
- iv) Stochasticity associated with sampling.

If too few years of data are collected it can be difficult to separate population trends from these other sources of underlying environmental stochasticity.

There are many analysis techniques for trend detection, some of which may be discussed later this afternoon. The appropriate analysis technique should be chosen prior to sampling to ensure that sufficient data is collected, and that the sampling design meets the assumptions of the analysis technique. For example, many techniques assume the samples are independent. Other techniques assume the sites are independent but require repeated samples to be taken from the same site over time.

Independence of data collected over time is often listed as a requirement, but in fact this statement should be considered more closely. Observations collected through time that are closely spaced can "carry over" information from one observation to the next. This is known as serial correlation, or, autocorrelation (Loftis *et al.*, 1991). However, the concept of "closely spaced" observations implies that a temporal scale has been defined.

The example given in Loftis *et al.* is firstly about the spatial scale of observations. Four 100ml samples drawn from one 5 litre bucket of water taken from a lake can not be considered four independent samples from the lake, but could be considered four independent samples from the bucket. The difference here is scale - the lake is the population of interest and so four 100ml samples from one bucket of water are not representative of the entire lake. When the bucket of water is the population of interest the 100ml samples could well be representative of the population.

Considering now temporal scale, seven daily observations within a week could not be considered independent samples for assessment of the water quality in a year, but could be if the study was interested in the water quality for the week.

The temporal scale of the study needs to be well understood, and defined. If a sequence of declining values in a measure of environmental quality was observed is this a "trend" or is it "autocorrelation"? A short-term decline could be seen as an artifact of autocorrelation if after a suitable time the measure returned to the overall average. But if interest is in the specific period of time of the short-term decline then it would be considered a trend (Loftis *et al.* 1991). The distinction between trend and autocorrelation is a function of the scale of interest.

2.4 Spatial Replication of Sites - Purposely Chosen or Randomly Chosen Monitoring Sites

For practical reasons often long-term monitoring sites are not randomly chosen. For example, the nine sites of the United Kingdom Environmental Change Network (ECN) were chosen on the basis of:

- i) Good geographical distribution covering a wide range of environmental conditions and the principal natural and managed ecosystems
- ii) Some guarantee of long-term physical and financial security
- iii) A known history of consistent management reliable and accessible records of past data, preferably for ten or more years; and,

- iv) Sufficient size to allow the opportunity for further experiments and observations.

The interest in the ECN is in monitoring the change in these sites and therefore it does not matter that the sites were not initially all similar in their status. The ECN is attempting to relate the differences in the change in sites to measured meteorological and geographical differences.

When selection is not based on randomness and probability it should be done cautiously to ensure sites are representative, and not e.g., sites with high initial animal abundance on productive sites. While it may seem sensible to choose sites where there are many animals, the changes in abundance these sites undergo may not be representative of the changes sites of lower productivity undergo. As another example, one type of non-probability sampling is called convenience sampling. Sites are selected based on their access, convenience, and the cost of sampling. The concern here is that the shift between sampled population and the target population may be large. Inference should be restricted to the population being sampled. When results are inferred to apply to the whole target population this will introduce bias and the variance will be underestimated (Van der Meer 1997).

The alternative design to purposely choosing sites is to randomly select sites. A random selection of sites ensures there is no bias in the estimation of population parameters. This attribute of random selection is discussed in many standard texts on sampling (e.g., Cochran 1977). In the strictest form random sampling is rarely used and other forms of probability sampling that should be considered include stratified sampling, cluster sampling, systematic sampling, and unequal probability sampling (two helpful texts on these and other designs are Thompson 1992, Lohr 2000).

2.5 Spatial Replication of Sites - Autocorrelation

Spatial autocorrelation is a much talked about issue in ecology and environmental science at the moment, partly because of the interest in geostatistical techniques. Here the interest is in the effect of autocorrelation on sampling design. Random variables are defined as being spatially autocorrelated if there is a tendency for pairs of variables that are a certain distances apart to be correlated. In other words, the values at one location can be predicted to some extent, by the values at neighbouring locations. This means the values are not stochastically independent (Legendre 1993).

Simple random sampling is a design-based approach where every unit has the same probability of being selected in a sample, and every sample, of size n , has the same probability of being selected. The randomization in the design ensures statistical independence of the n random variables regardless of spatial autocorrelation (Aubrey and Debouzie 2000). The variables are random because their locations are random and have been independently selected (De Gruijter and Ter Braak 1990). This fact is ensured by the design. Having random variables does not mean that there is not spatial autocorrelation (if it exists) in the data, but that the estimation of the variance of the mean (i.e., σ^2/n) is unbiased (Brus and de Gruijter 1993).

Bias in the estimated variance of the mean can be introduced when the simple random formulae is used for systematic sampling because of the dependence of the data and the spatial dependence of the population (Aubrey and Debouzie 2000). When the underlying population is random (i.e., no spatial autocorrelation), there should not be bias, but in other situations the estimation of the variance of the mean will be biased (either overestimated for populations that have an underlying trend, or underestimated for populations that have a systematic pattern that matches the systematic sampling interval).

There are a number of other approaches that can be used for estimation of variances for systematic sampling. Design-based methods include multi-start systematic samples (Gautschi 1957), methods where the adjacent points are grouped and the sample treated as a stratified

sample (Yates 1961), and serpentine patterns. There are also a number of model-based approaches that specifically account for autocorrelation. Generally these model-based approaches are geostatistical methods. An interesting observation is that while these estimators have been published, they are rarely used by field-biologists (Aubrey and Debouzie 2000). This may be because they added an extra layer of complexity to sampling programmes.

A geostatistical model-based approach starts with the assumption that there is a model that generates random variables over the space A (De Gruijter and Ter Braak 1990). The sample values are from one population from an infinite set of possible populations generated from the model (Brus and de Gruijter 1993). Sampling fixed points means that their locations are not random even if the measurements taken at each location are considered random. This is best explained with some notation (De Gruijter and Ter Braak 1990). Within a region A , z is measured at n points. The values are $z(x_i)$ where $i = 1, \dots, n$, and x_i is the location of the sample point. In the design-based approach such as simple random sampling, the n sample points are randomly chosen and the variables are $z(X_i)$ where capital letter means random and small letter means nonrandom. In a model-based approach the variables are $Z(x_i)$. In this framework, $z(X_i)$ and $z(X_j)$ can be considered stochastically independent while $Z(x_i)$ and $Z(x_j)$ can not be (although this depends on the model that was used to generate the random variables over the space).

There is debate on the idea that classical inference can be used on data from random sampling when there is underlying spatial structure. Some authors, e.g., Fortin *et al.* (1989), Legendre (1993) argue that the observations are dependent hence standard statistical hypothesis tests can not be used. However, *classical sampling theory* should be distinguished from *classical statistical inference*. Classical sampling theory allows inference about the population based on the probability of selecting a given set of sample units, or locations. This probability is defined by the sample design. The value of z is measured without error - it is nonrandom. In contrast, in model-based inference, the values at each location are considered random and the inference is based on the procedure of defining the model (De Gruijter and Ter Braak 1990).

2.6 Spatial Replication of Sites - Resample or Reselect?

Even with randomly selected sites there is still the question of what to monitor over time - do you measure the same sites at each time period, or randomly reselect sites at time period? The answer depends on the sampling objective - is it to determine status or trend? Status refers to questions like, "How many possums are there?", "What is the water quality of the river?", questions about extent, productivity and condition. Trend refers to changes in these with time (Olsen *et al.* 1999).

- If the objective is to estimate the mean value following the most recent survey, e.g., environmental status, then it is best to reselect a fresh sample (i.e., new sample locations).
- If the objective is to estimate the change in population means, i.e., trends in environmental status it is best to use the same sites for each survey, (Skalski 1990).

With the former case, by reselecting the sites each year the population parameter will not be consistently over- or under-estimated. With the later case, resampling the same site each year will eliminate random variation among sites that could confound the survey results.

In the framework of univariate analysis of variance, randomly selecting b sites in each of the A years can be modeled as a two factor nested model with the random factor sites nested within years (fixed factor). When the same sites are resurveyed each year after the initial random selection of sites the model is a two factor mixed model with the random factor sites and fixed factor years. Finally, the design above where sites are not randomly chosen and are sampled each year is a simple two factor model with both sites and years fixed. With this third design, it would be misleading to draw inference to the whole study area and would lead to bias (Van der Meer 1997).

If the among site variance is large, and all sites have much the same annual trend, then revisiting the same sites each year will be a more powerful design for detecting among year change than randomly reselecting sites each year. Another way of saying this is that when there is, on average, a positive correlation among the means of the sites between two successive years revisiting the same sites each year is more powerful. A more technical derivation of this result is given in Van der Meer (1997). There are many other models that can be used here and some of these would be more appropriate when the assumptions of the analysis of variance are not met (see summary in Van der Meer 1997).

2.7 Some Special Designs for Choosing Monitoring Sites

Monitoring can often have both objectives described above - to detect status as well as to detect trends. Skalski (1990) suggested an *augmented rotating panel design* for long-term monitoring. This design combines both ideas where some sites are sampled every year and others are rotated.

The design takes the form shown in Table 2.1 if there are eight sites that are visited every year and four sets of ten sites that are rotated. Site set 7, for example, consists of ten sites that are visited in years 4 to 6 of the study. The number of sites in different sets is arbitrary. Preferably, the sites will be randomly chosen from an appropriate population of sites.

This design has some appealing properties: the sites that are always measured can be used to detect long-term trends but the rotation of blocks of ten sites ensures that the study is not too dependent on an initial choice of sites that may be unusual in some respects. However, Urquart *et al.* (1993) have provided evidence that the serially alternating design that is discussed next is more efficient because more sites are measured in the first few years of the study.

The practical reasoning behind the rotating panel design is very sensible. When there is limited information on an environmental impact it is difficult to know where and when to monitor. Without knowing the extent of the area effected by e.g., a coastal sewer out-fall the monitoring design should include many sites that are situation from the out-fall in all directions (land and sea) and for a good distance. Such large spatial coverage is often beyond the scope of most budgets. This design provides a method to "add in" new sites and allows for the dynamic nature of populations (human and non-human). For example, monitoring of an out-fall in Akaroa Harbour may focus on the areas where people live. Over time the new housing developments may mean that new monitoring sites should be added into the survey design.

Table 2.1 Augmented rotating panel design. In this example, every year 48 sites are visited. Of these, 8 are always the same and the other 40 sites are in four block of size ten, such that each block of ten remains in the sample for four years after the initial start up period.

Site set	Number of sites	Years											
		1	2	3	4	5	6	7	8	9	10	11	12
repeated	8	x	x	x	x	x	x	X	X	x	x	x	x
1	10	x											
2	10	x	x										
3	10	x	x	x									
4	10	x	x	x	x								
5	10		x	x	x	x							
6	10			x	x	x	x						
7	10				x	x	x						
..													
14	10											x	x
15	10												x

The *augmented serially alternating design* is similar to the design discussed above but rather than surveying 30 out of 40 sites in the rotating panel the next year, each year a rotating selection of 40 sites are surveyed (Table 2.2). The US Environmental Protection Agency Environmental Monitoring and Assessment Program (EMAP) was based on this design.

Table 2.2 Augmented serially alternating design. In this example, every year 48 sites are visited. Of these, eight are always the same and the other 40 sites are from a rotating panel.

Site set	Number of sites	Years											
		1	2	3	4	5	6	7	8	9	10	11	12
repeated	8	x	x	x	x	x	x	x	x	x	x	x	x
1	40	x				x				x			
2	40		x				x				x		
3	40			x				x				x	
4	40				x				x				x

The advantage of this serially alternating design is that more sites are visited. Compare Table 2.1 and 2.2. At the end of the first year 48 sites will have been surveyed. By the end of year 2 with the serially alternating design 88 sites will have been visited, but with the design shown in Table 2.1 only 58 sites will have been visited. By the end of year 3, 128 will have been visited with the serially alternating design and only 68 with the other design, and so on. The point is that with the serially alternating design more sites are visited, and more information is collected from among the sites. This is discussed further in the next section on power, but generally it is better to collect information from as many sites as possible. With both designs eventually all 168 sites will be surveyed. With the serially alternating design this will take 4 years, but with the other design it will take 13 years.

Methods for analysis of data from these designs for trend and status are discussed in Skalski (1990) and Urquart *et al.* (1993).

2.8 Statistical Power

One of the major considerations in designing a monitoring scheme is whether you will be able to detect a true change, or trend, in the population parameter of interest (Taylor and Gerrodette, 1993, Fairweather 1991). This ability to detect a trend, e.g., the effect of human-induced change, which occurs over and above the amount of variation that natural populations exhibit is referred to as power.

When planning a monitoring study the number of samples, the likely effect that can be detected, and the number of years required to be able to detect a trend are all considerations that relate to power. Calculating the power of a trend survey can be difficult because it requires estimates of variance and until monitoring is undertaken there may not be any estimates of the variances (Gerrodette 1987, Steidl *et al.* 1997). However, approximations of likely power can be made by using data from other studies and from pilot studies.

In statistical terms the power of a test is the ability to reject the null hypothesis when it is false, that is, it is the probability of correctly rejecting the null hypothesis. In trend detection power refers to the ability to detect a true increasing or decreasing trend. Several factors affect power, such as sample size, variability of the samples, and magnitude of the difference or trend to be detected. Designs with small sample sizes and high variability will have low power. If the size of the difference or trend is small compared with the natural population variability it will be difficult to detect any effect.

Typically type I errors are set at 0.05 ($\alpha = 0.05$). In setting such a low type I error rate there is an implicit assumption that the relative cost of a type I error is higher than a type II error (β) because in most situations as α decreases β increases (Steidl *et al.* 1997). In designing a monitoring programme the evaluation of the relative costs of type I and type II errors should be made. Often the cost of a type II error far outweighs the cost of a type I error - the cost of failing to recognise a detrimental effect on the environment against falsely stating there has been an effect (Petermen 1990, Fairweather 1991). However, in some situations the relative costs are reversed, for example, in pest control failing to detect a reduction in the population after a control operation is of less concern than falsely detecting a reduction when the control operation was unsuccessful (Brown and Miller 1998). Mapstone (1995) suggests a method to set type I and II errors based on the relative costs of the two errors. Essentially the relative cost is determined (more correctly estimated) e.g., $R_c = \text{cost of type II} / \text{cost of type I}$. The size of the type II error is then determined as α / R_c .

2.10 How to Improve Power

Variation in data collected from monitoring studies is due to trend from the population chance and from:

- i) Spatial variation (variation among sites due to the environmental heterogeneity),
- ii) Temporal variation (variation over time at the scale of interest), and
- iii) Within-site variation (variation due to inexactness of the data collection).

(Millard and Lettenmaier 1986, Gerrodette 1987, Link *et al.* 1994). The power of monitoring, for example, the ability to detect if there is a true difference between a treatment and a non-treatment site, or to detect a regional-population trend, will improve if these sources of variation are reduced.

Within-site variation can be considered measurement error, or small-scale spatial and temporal variation. What is measurement error and what is spatial or temporal variation is therefore related directly to the survey objective and the definition of the sampling unit. Variation within the sampling unit, where the sampling unit is the site, is within-site variation, or measurement error. Variation within months would be within-site variation, or measurement error when the interest is in detecting annual trends. It is important not to ignore within-site variation, but to design the survey so that the within-site variation is less than the among site variation, and so that small-scale temporal variation is less than the larger scale variation that is of interest. If this is not considered in the design then an observed variation among years could be due to small-scale fluctuations within each year and not a long-term trend. Equally the observed variation among years (temporal variation) could be due to small-scale spatial variation (Morrissey *et al.* 1992). Such confounding is best controlled by sample design - the size of sample units and replication.

Strategies to reduce within-site variation are related to survey design, e.g., to have strict guidelines of when and how sampling should be undertaken, and taking replicate samples (although these can not be considered true independent replicates). Modelling the environmental factors that effect the observed sample values can also reduce within-site variation. For example, consider the example described above for monitoring Karner blue butterflies in Wisconsin. Butterflies are less mobile and less detectable on cool days compared with warm days. Counts of butterflies seen during surveys (i.e., the observed sample data) on cool days can be inflated to adjust for differences in daily temperature.

Strategies to reduce spatial and temporal variation depend on the relative size, and scale, of the spatial and temporal variation and on the scale of the monitoring e.g., to detect long term annual trends or short term trends over a few years. Survey effort can be allocated into surveying more sites, putting more effort into a site, or surveying more frequently within a season or among seasons in the year. The best approach in designing a survey is to collect data and

estimate power. However, there are some general trends that emerge from the literature on these sorts of studies.

For trend detection sampling more units is generally preferable to increasing sampling effort within a unit (Millard and Lettenmaier 1986, Link *et al.* 1994, Lester *et al.* 1996, Brown and Miller 1998). Millard and Lettenmaier (1986) found that in their study to maximise power the optimal design was a spatially extensive one with many sampling units. With a design with many sample sites the among-site variation is reduced.

Similarly, survey effort can be increased by resampling the same site (e.g., Kendall *et al.* 1992, Lessica and Steele 1996). However, there will be a limit to how much gain in the ability to detect a trend can be made by sampling more frequently within a season when the trend is being measured over years (Lester *et al.* 1996). Sampling more frequently within a season runs the risk that the data are serially correlated. If this is the case then the subsequent measurements are not true replicates and what may have been achieved is more "within-site" effort as discussed above. In a study of beach seine surveys sampling the same site twice was found not to be useful because the second survey occurred sooner after the first, for logistical reasons, and the two surveys could not be considered true replicates. The second survey always had lower counts than the first suggesting fish were removed in the first survey or were scared by it. The optimal design was to survey more beaches (i.e., increase spatial replication) and to use a slightly larger net (i.e., reduce within site variation) (Wilson and Weisburg 1993).

When the site is resurveyed some alternatives for analysis, other than considering the resurveys as increased within-site effort, are to ignore serial correlation or avoid serial correlation. When serial correlation is ignored this will cause the variance to be underestimated and increase the type I error rate. Serial correlation can be avoided by spacing surveys further apart in time. This later option reduces power because there are fewer sample points (Schroeter *et al.* 1993).

So far we have been discussing designs for long-term monitoring to detect trends in population status. Monitoring is also undertaken to detect a possible change following some specific management action, e.g., to detect if a site has been cleaned up after a remedial action, or to detect whether a rat population has been reduced after a rat-poisoning operation. In these situations one way to improve power is to have "treatment" and "non-treatment" sites.

With "treatment" and "non-treatment" sites the differences between the sites can be compared over time. The power to detect the difference between trends in treatment and non-treatment sites will generally be higher than the power to detect the individual trend at either site. If the variation among time intervals for the treatment sites was identical to the variation for the control sites, by using the differences between the two, this source of variation would be eliminated (Stewart-Oaten *et al.* 1986). Even if the correlation is not perfect, Stewart-Oaten argue that the variation in the differences over time would be small compared to other sources of variation, particularly from sampling error.

2.11 What Sample Unit to Use

The decisions on what sampler to use is usually defined by the population, that is the physical characteristics of the habitat and the type of organisms (Resh 1979). For example, in surveys of low growing weeds a square plot (a quadrat) is often used. In freshwater fisheries surveys where electric fishing is used the sample unit is a site, measured in metres. Air samples may be units that are a volume of air collected over a time period. In marine fisheries surveys trawls may be the sample unit. In the case study below the sampler is a cylindrical core - either of 13 cm in diameter or twice that size. The optimal size is to be determined.

The actual sampling device often determines the size of the sampler, for example, a trawl net is a fixed size and although the length of the tow may vary, smaller, or larger nets may not be

feasible. However, plots, are not a fixed physical unit and can be e.g., 1m^2 , or 0.25m^2 , or 10m^2 etc. depending on the population of interest. One rule of thumb for plot size is that the plot should be 20 times the size of the individual in the population (Green 1979).

In general, the larger the sample size the better so smaller sample units can be advantageous. For surveys that use plots many, small plots rather than a few large, plots is recommended. However there must be a balance between the size of the sample and the size of the unit within the sample. For example, if the plot is so small that it is the same size as the individuals in the population the sample will be highly variable because it will consist of either plots of zero counts and plots with counts of one. At the other extreme, if the plots are very large the variability among sample units will be low, but there will be few plots.

Another general rule for sample unit selection is that the size of the unit should not match the scale of any patchiness in the environment. For example, if the plots used to sample the low growing weed were 1m^2 and the weeds occurred in patches that were about 1m^2 in size then some of the plots would have very high counts (when the plot was located entirely within a patch), and others would have very low counts (when the plot was located entirely outside a patch). The sample would have high variance, and low precision. The plot should either be very much larger than the scale of patchiness, or very much smaller.

One other consideration in using plots is their shape. A long, thin rectangular plot is an efficient shape because the plot "spreads" across more of the study area. This has the effect of minimising correlation between individuals within the plot and therefore the plot is more informative. A circular plot has less spatial "spread" and higher correlation within the plot. The advantage of a circular plot is that it will have less edge than a long, thin rectangular plot. The problem with a plot shape with a lot of edge is there is more chance of mistakenly recording an individual as "in" the plot when in fact it is "out" and vice versa.

Often the best solution to what size sample unit to use is to conduct a pilot survey with various sample unit sizes to give information on the precision and total sample cost of each (Green 1979). Practical considerations need to be taken into account, and some these may not be obvious until a field trial. For example, in a pilot study to assess various quadrat sizes for surveys of aquatic macrophyte by divers using SCUBA, the smaller quadrats were too light and tended to slide off the plant clumps onto bare ground leading to underestimates of abundance (Downing and Anderson 1985). Other methods are to use a nested design in the pilot survey where many small units are used. Estimates of precision and cost are calculated using the smallest sample unit size. Then, adjacent units are combined to give an effective sample unit size that is twice as large and new estimates of precision and cost are calculated. These units can be combined again, and so on, giving a range of successively larger sample units and estimates of precision and cost can be compared between the various sizes.

2.12 Errors in Sample Surveys

In general there are four sources of error or variation in scientific studies (Cochran, 1977):

- i) There are sampling errors due to the variability between experimental units and the random selection of units included in a sample. Different random samples will generally produce different estimates of population parameters. This variation reflects the sampling errors.
- ii) There may be measurement errors due to the lack of uniformity in the manner in which a study is conducted. The measurement procedure may be biased, imprecise or both biased and imprecise. This type of error results solely from the manner in which the observations are made. For example, fisherman may report incorrect lengths and weights of fish caught, human subjects may lie about their age or weight, etc.
- iii) There may be missing data due to the failure to measure some units in the sample.
- iv) Gross errors may be introduced in coding, tabulating, typing and editing data.

An understanding of sampling errors and their effects is the basis of statistical inference procedures. The control of sampling errors is therefore primarily the responsibility of the statistician. Random measurement errors can be modelled but their control and reduction must come from careful experimental design. In fact, in many fields of study the presence of measurement error is barely recognised and its influence is played down. Many statisticians follow the rule of thumb that the measurement error should be small relative to the sampling error, especially in utilising statistical procedures such as regression and correlation analysis. Certainly for many studies conducted in ecology measurement errors cannot be ignored and standard analysis procedures such as regression analysis may not be applicable until this source of error is under control.

References

- Aubrey, P. and Debouzie, D. (2000). Geostatistical estimation variance for the spatial mean in two-dimensional systematic sampling. *Ecology* 81:543-553.
- Barker, R.J. and Sauer, J.R. (1992). Modelling population change from time series data. In: *Wildlife 2001: Populations*. eds. McCullogh, D.R. and Barrett, R.H. Elsevier Applied Science, London.
- Brown, J.A. and Miller, C.M. (1998). Monitoring stoat *Mustela erminea* control operations: power analysis and design. *Science for Conservation*: 96, Department of Conservation, Wellington
- Brown, J.A. and Boyce, M.S. (1996). Monitoring of Karner blue butterflies (*Lycaeides melissa samuelis*) for the proposed habitat conservation plan, Wisconsin. Report to the US National Fish and Wildlife Foundation.
- Brus, D.J. and de Gruijter, J.J. (1993). Design-based versus model-based estimates of spatial means: theory and application in environmental soil science. *Environmetrics* 4(2):123-152.
- Caselton, W.F., Kan, L. and Zidek, J.V. (1992). Quality data networks that minimize entropy. In *Statistics in the Environmental and Earth Sciences* (eds. A.T. Walden and P. Guttorp), pp. 10-38. Edward Arnold, London.
- Cochran, W.G. (1977). *Sampling Techniques*. 3rd edition. Wiley, New York.
- De Gruijter, J.J. and Ter Braak, C.J.F. (1990). Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematics Geology* 22:407-415.
- Downing, J.A. and Anderson, M.R. (1985). Estimating the standing biomass of aquatic macrophytes. *Canadian Journal of Fisheries and Aquatic Sciences* 42:1860-1869.
- Eberhardt, L.L. (1976). Quantitative ecology and impact assessment. *Journal of Environmental Management* 4: 27-70.
- Elinzga, C.L.D., Salzer, D.W. and Willoughby, J.W. (1998). Measuring and monitoring plant populations. Bureau of Land Management Technical Reference 1730-1.
- Fairweather, P.G. (1991). Statistical power and design requirements for environmental monitoring. *Australian Journal of Marine and Freshwater Research* 42: 555-567.
- Fedorov, V. and Mueller, W. (1989). Comparison of two approaches in the optimal design of an observation network. *Statistics* 20: 339-51.

- Fortin, M., Drapeau, P. and Legendre, P. (1989). Spatial autocorrelation and sampling design in plant ecology. *Vegetatio* 83:209-222.
- Gautschi, W. (1957). Some remarks on systematic sampling. *Annals of Mathematical Statistics* 28:385-394.
- Gerrodette, T. (1987). A power analysis for detecting trends. *Ecology* 68: 1364-1372.
- Gibbs, J.P., Snell, H.L. and Causton, C.E. (1999). Effective monitoring for adaptive wildlife management: Lessons from the Galápagos Islands. *Journal of Wildlife Management* 63:1055-1065.
- Green, R.H. (1979). *Sampling Design and Statistical Methods for Environmental Biologists*. Wiley, New York.
- Kendall, K.C., Metzgar, L.H., Patterson, D.A., and Steele, B.M. Power of sign surveys to monitor population trends. *Ecological Applications* 24(2):422-430.
- Legendre, P. (1990). Spatial autocorrelations: Trouble or new paradigm? *Ecology* 74(6).
- Lessica, P. and Steele, B.M. (1996). A method for monitoring long-term population trends: an example using rare arctic-alpine plants. *Ecological Applications* 6(3):879-887.
- Lester, N.P., Dunlop, W.I. and Willox, C.C. (1996) Detecting changes in the nearshore fish community. *Canadian Journal of Fisheries and Aquatic Sciences* 53(Suppl. 1):391-402.
- Link, W.A., Barker, R.J., Sauer, J.R. and Droege, S. (1994). Within-site variability in surveys of wildlife populations. *Ecology* 75:1097-1108.
- Loftis, J.C., McBride, G.B. and Ellis, J.C. (1991). Considerations of scale in water quality monitoring and data analysis. *Water Resources Bulletin* 27:255-264.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury, Pacific Grove.
- Mapstone, B.D (1995). Scaleable decision rules for environmental impact studies: effect size, type I, and type II errors. *Ecological Applications* 5:401-410.
- Millard, S.P and Lettenmaier, D.P. (1986). Optimal design of biological sampling programs using analysis of variance. *Estuarine, Coastal and Shelf Science* 22:637-656.
- Morrissey, D.J., Underwood, A.J., Howitt, L., and Stark, J.S. (1992). Temporal variation in soft-sediment benthos. *Journal of Experimental Marine Biology and Ecology* 164:233-245.
- Olsen, A.R., Sedransk, J., Edwards, D., Gotway, C.A., Liggett, W., Rathbun, S., Reckhow, K.H. and Young, L.J. (1999). Statistical issues for monitoring ecological and natural resources in the United States. *Environmental Monitoring and Assessment* 54: 1-45.
- Overton, W.S., White, D. and Stevens, D.L.. (1991). *Design Report for EMAP, the Environmental Monitoring and Assessment Program*. U.S. Environmental Protection Agency report EPA/600/3-91/053, Washington, D.C.
- Peterman, R.M. (1990). Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences* 47:2-15.

- Resh, V.H. (1979) Sampling variability, life history features, and the experimental design of aquatic insect studies. *Journal of the Fisheries Research Board of Canada* 36:290-311.
- Ringold, P.L., Alegria, J., Czapski, R.L., Mulder, B.S., Tolle, T. and Burnett, K. (1996). Adaptive monitoring design for ecosystem management. *Ecological Applications* 6:745-747.
- Schroeter, S.C., Dixon J.D., Kasendiek, J., Smith, R.O., bence, J.R. (1993). Detecting the ecological effects of environmental impacts: a case study of kelp forest invertebrates. *Ecological Applications* 3(2):331-350.
- Skalski, J.R. (1990). A design for long term status and trends monitoring. *Journal of Environmental Management* 30:139-144.
- Skalski, J.R. and McKenzie, D.H. (1982). A design for aquatic monitoring programs. *Journal of Environmental Management* 14:237-251.
- Steidl, R.J., Hayes, J.P., Schaubert, E. (1997). Statistical power analysis in wildlife research. *Journal of Wildlife Management* 61:270-279.
- Stevens, D.L. and Olsen, A.R. (1991). Statistical issues in environmental monitoring and assessment. In: *Proceedings of the Section on Statistics and the Environment*, American Statistical Association, Alexandria, Virginia.
- Stewart-Oaten, A., Murdoch, W.W. and Parker, K.R. (1986). Environmental impact assessment: "Pseudoreplication" in time? *Ecology* 67:929-940.
- Taylor, B.L., Gerrodette, T. (1993). The uses of statistical power in conservation biology: the vaquita and northern spotted owl. *Conservation Biology* 7:489-500.
- Thompson, S.K. (1992) *Sampling*, Wiley.
- Underwood, A.J. (1992). Beyond BACI: The detection of environmental impacts in the real, but variable, world. *Journal of Experimental and Marine Biological Ecology* 161:145-178.
- Underwood, A.J. (1994). On beyond BACI: Sampling designs that might reliably detect environmental disturbances. *Ecological Applications* 4:3-15.
- Underwood, A.J. (1997). *Experiments in Ecology*. Cambridge University Press, Cambridge.
- Urquhart, N.S., Overton, W.S. and Birkes, D.S. (1993). Comparing sampling designs for monitoring ecological status and trends: impact of temporal patterns. In: *Statistics for the Environment*, eds. Barnett, V. and Turkman, K.F. Wiley and Sons.
- Van der Meer, J. (1997). Sampling design of monitoring programmes for marine benthos: a comparison between the use of fixed versus randomly selected stations. *Journal of Sea Research* 37:167-179.
- Wilson, H.T. and Weisburg, S.B. (1993). Design considerations of beach seine surveys of striped bass. *North American Journal of Fisheries Management* 13:376-382.
- Yates, F. (1961). *Sampling Methods for Censuses and Surveys*, 4th ed., Oxford University Press.

A Case Study: Monitoring Estuarine Quality

Stephanie Turner
Environment Waikato, Po Box 4010, Hamilton East, New Zealand
Jennifer Brown
University of Canterbury

3.1 Introduction

Estuaries are an important feature in the coastal environment of the Waikato Region, New Zealand. The Regional Authority, Environment Waikato, have selected estuaries as the main focus of their coastal ecological monitoring programme. The aim of the estuarine component of the programme is to determine current status and monitor temporal changes in the "health" of the Region's estuaries. Two key parameters will be included in the monitoring programme:

- i) Broad-scale surveys and monitoring of the extent and condition of important estuarine habitats,
- ii) Monitoring of intertidal soft-sediment benthic communities in estuaries.

This case study deals with only the second of these - monitoring intertidal estuarine soft-sediment benthic infaunal communities.

Intertidal soft sediments are a major component of estuaries in the Waikato Region. In a number of estuaries they represent more than ½ of the total area. Monitoring the key benthic infauna will give information on the "health" of the Region's estuaries. Benthic infauna are effective indicators of estuarine health - they are an important component of the ecosystem e.g., they are prey species for various fish and birds, they are important in the process of transferring organic matter, nutrients etc, and they are important in the stabilization and reworking of sediments.

The monitoring programme is not finalized at this stage, but the discussion that follows outlines some of the main features of it.

3.2 Estuary selection

There are roughly 35 estuaries in the Waikato Region. Not all are included in the monitoring programme. Rather, estuaries were grouped according to physical attributes (e.g., geomorphologic and oceanographic characteristics). The idea here is that within each group the environmental processes are likely to be more similar and which are likely to respond to stresses in a similar manner. Estuaries that were considered "representative" were selected from each group. Other issues that were considered in the selection were:

- i) The current issues and potential impacts (e.g., marine farms, marinas, reclamation),
- ii) Existing information and current/proposed research,
- iii) Community interest and support,
- iv) Logistical issues (distance to travel to the estuaries, access etc).

In total four estuaries have been selected – southern Firth of Thames, Coromandel Harbour, Tairua Harbour, and Whaingaroa Harbour.

3.3 Sample sites

Within each estuary there are a number of sites. Sites are distributed throughout the estuary by a stratified design. A preliminary visit to each estuary was made to visually assess areas of appropriate habitat. Then, the appropriate habitat was stratified by location e.g., entrance, mid-, upper-harbour and sites were randomly located within each stratum. At least two sites were selected in each stratum. A grid, laid over the appropriate habitat was used to randomly choose the sites. This stratified design ensured some spatial spread of sites while retaining the element of randomisation in site selection.

Sites are permanently marked by pegs and by GPS location and are about 0.1 ha in size. At each site randomly located cores will be collected and the numbers of individual species in each core counted. A stratified design will be used to ensure spatial coverage of the site by dividing each site into equal-sized sectors. This is largely based on the design developed by NIWA for the Auckland Regional Council and for a client in Whangapoua (Thrush *et al.*, 1988).

The number of sectors will equal the number of cores. The cores will be randomly located within the sectors, i.e., one core per sector, at each sampling event. If a core location is within 0.5m of the location of the core in the previous sampling event the location will be selected again. This follows the recommendation of Thrush *et al.* (1988) to preclude any modification of the sediment or resident populations from the previous sampling occasion. Also, cores will be at least 5m apart from cores in adjacent sectors to ensure the desired spatial coverage (Thrush *et al.* 1998).

Sample cores will be 13 cm diameter and approximately 15-20 cm deep. This is consistent with other estuary monitoring programmes elsewhere and will assist in any comparison among programmes. The number of cores is being determined by data from a pilot study.

3.4 What species to monitor

One of the considerations to make in designing the programme was what species to monitor. Various options were considered including:

- i) Monitoring one or two "indicator" species that are selected on the basis that they are representative of change in the whole community,
- ii) Monitoring a suite of species that are most susceptible to change,
- iii) Monitoring a suite of species selected on the basis of them being ecologically important, representative of change in the whole community, responding in characteristic manner to particular disturbances, and including a variety of taxonomic groups,
- iv) Monitoring all species to provide information on changes in abundance, and diversity.

The third option was chosen and 20 - 30 species are being identified and counted. In addition all other organisms are being identified to the level of major taxonomic groups to provide some information about change in overall community biodiversity.

3.5 Survey design

A simple approach to estimate optimal sample size is to consider the reduction in standard error (*se*) with increasing *n*. The *se* will decrease monotonically with increasing *n*, and the minimum sample size should be beyond the region of maximum change. There are various ways to do this using slightly more sophistication, for example, using the maximum, or 95th percentile of the standard errors calculated from all the possible samples of size *n* (Bros and Cowell 1987, Hewitt *et al.* 1993).

Design of the survey is slightly more complicated in this study because there are two spatial scales - sites and cores within sites. An assessment of likely power was conducted using a randomisation programme, MONITOR. It is important to distinguish between actual statistical power and this "prospective" power analysis where the analysis is carried out to help decide in the actual design (Thomas 1997). The importance of this analysis is not in determining the actual power but more in understanding what drives power and what power actually is. In this study the statistical computations used in the computer package do not mimic the statistical techniques that will be used to analysis the data from the monitoring programme. The package was used because it is easy, gives results quickly, and does not rely on complicated input data. In this study, the power analysis was to help define the sample design. Other considerations were taken into account in defining the design, including practical considerations. For example, how many sites and cores can be sampled in one day. There would be little gained if the theoretical optimal

number of sites that should be visited in an estuary took 1.2 days because of the additional cost of returning to the estuary on a second day to collect a few cores.

The preliminary power analysis for the annual sampling suggested that the best design was one where there were many cores (20 - 30) and fewer sites (4 - 8). This was based on annual samples and at that stage in the design process the recommended approach was to collect 20 cores from 8 sites (Figure 3.1).

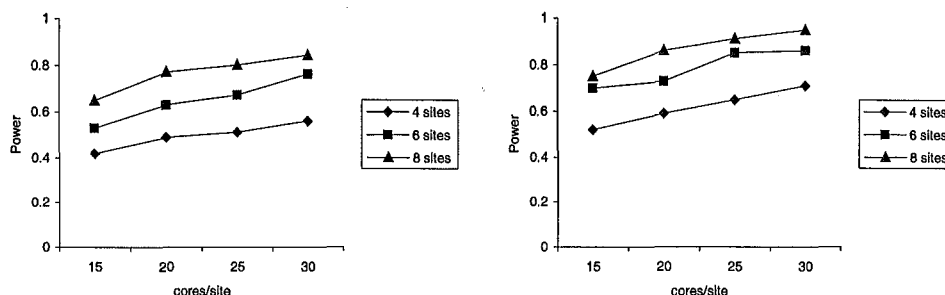


Figure 3.1. Estimates of likely power to detect a -3% and +3% annual change, over 10 years, for *Aquilaspio* sp.(a species of polychaete or marine worm) for various combinations of sites/estuary and cores/site sampled once a year, and with $\alpha = 0.20$.

The detection of long-term trend can be complicated by the presence of within-year variability. The estuaries have a number of temporal-cycles that occur at different scales e.g., daily tidal cycles, monthly cycles that affect the tide heights, and seasonal variation. Estuarine benthic macrofauna can exhibit considerable within year variability. One strategy to deal with within-year variation is to sample at the same season in every year, e.g., the same month (Olsen *et al.* 1999, Alden *et al.* 1997). This limits the extent of the monitoring to that season of the year unless the same trend occurs in all seasons. For this study the decision was made to incorporate within-year sampling to give information on changes at a smaller than annual scale.

The estimated power improved by using a twice-yearly sampling regime (Figure 3.2). Now the best design appears to be one with even fewer sites, 3, with 12 cores. Twice-yearly sampling would be once in April and once in October.

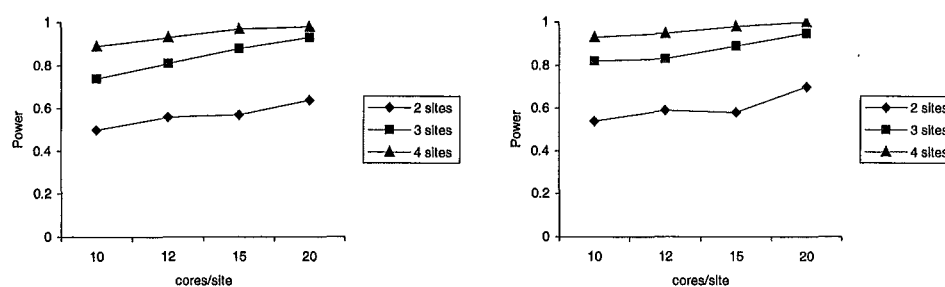


Figure 3.2. Estimates of likely power to detect a -3% and +3% annual change, over 10 years, for *Aquilaspio* sp.(a species of polychaete or marine worm) for various combinations of sites/estuary and cores/site sampled twice a year, and with $\alpha = 0.20$.

Another strategy is to sample more frequently than twice-yearly, e.g., seasonally, or quarterly. However, with such frequent visits to the estuaries the additional gain in information

maybe marginal given the additional costs. In a similar programme in Chesapeake Bay while sampling in each season was the most powerful design, sampling in two seasons resulted in only a small loss in power (Alden *et al.* 1997). In Chesapeake Bay they observed the same trends within each season at all sampling sites, which intuitively suggests sampling in four seasons is not necessary. Sampling in two seasons is a compromise between one- and four-season sampling. For the same total cost, one-season sampling can be thought of as a way of maximizing spatial coverage and four-season sampling maximizing temporal coverage. Sampling more than once within a season will provide more information for characterizing that season but not necessarily improve trend detection.

As a final step in the development of the monitoring programme, some information was available from a pilot study. The final design chosen, at this stage, is to sample sites 6-monthly (April, October). There will be five sites in an estuary with 12 cores (Figure 3.3). In addition, two of the sites will be sampled more frequently at quarterly intervals (January, April, July, October). This quarterly sampling is a compromise between the gain in extra information and the extra costs. At the 6-year review the benefit from the limited quarterly sampling can be evaluated.

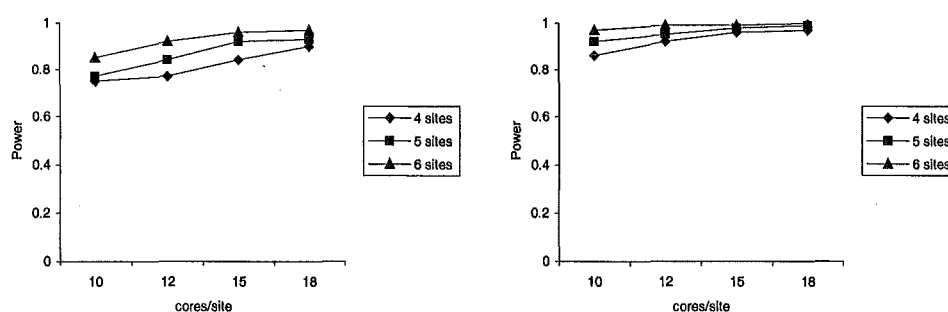


Figure 3.3. Estimates of likely power to detect a -3% and +3% annual change, over 10 years, for *Aquilaspio sp.* (a species of polychaete or marine worm) for various combinations of sites/estuary and cores/site sampled twice a year, and with $\alpha = 0.20$. Estimates based on preliminary pilot study data.

A rotating panel design will be used where sites are sampled for 6 years. Then, the site will be dropped from the programme and a new site added in.

3.6 Analysis

The actual analysis method will depend on the specific research objective but there are a number of different techniques that can be used. The data being collected are at two different spatial scales - within sites (cores) and among sites. More correctly there are three scales - among estuaries - but at this stage the focus is on individual estuaries.

As with any analysis the first step will be displaying the data in useful ways. The data collected, at each sampling event, can be considered multivariate. The data is a vector of counts of the individual species. Counts are either at the scale of cores, or the average count for cores within a site. Some examples of multivariate analyses that can be used are principal components plots, plots of clusters from a cluster analysis, and plots from multidimensional scaling. A nonparametric method to track changes over time in the plots from multidimensional scaling is called analysis of similarities (Clarke 1993) and can itself produce plots that are visually easy to interpret. In this technique a matrix of similarity indices are calculated from pairs of observations in the multivariate data (i.e., numbers of individuals of the indicator species, for each sampling event), and statistical significance is estimated from a randomisation test based on an assumption of no differences.

Analysis of trend can use some time-series analysis or other trend detection methods. The data has two temporal scales - annual (at all sites) and within-year (at a reduced, and changing, set of sites) - and this should allow some separation of within-year and among year variation.

References

- Alden, R.W., Weisburg, S.B., Ranasinghe, J.A. and Dauer, D.M. (1997). Optimizing temporal sampling strategies for benthic environmental monitoring programs. *Marine Pollution Bulletin* 34(11):913-922.
- Bros, W.E. and Cowell, B.C. (1987). A technique for optimizing sample size (replication). *Journal of Experimental marine Biology and Ecology* 114:63-71.
- Clarke, K.R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* 18:117-143.
- Hewitt, J.E., McBride, G.B., Pridmore, R.D., Thrush, S.F. (1993). Patchy distributions: optimizing sampling size. *Environmental Monitoring and Assessment* 27:95-105.
- Olsen, A.R., Sedransk, J., Edwards, D., Gotway, C.A., Liggett, W., Rathbun, S., Reckhow, K.H. and Young, L.J. (1999). Statistical issues for monitoring ecological and natural resources in the United States. *Environmental Monitoring and Assessment* 54: 1-45.
- Thomas, L. (1997). Retrospective power analysis. *Conservation Biology* 11:176-280.
- Thrush, S.F., Pridmore, R.D., Hewitt, J.E. and Roper, D.S. 1988. Design of an ecological monitoring programme for the Manukau Harbour. Report to the Auckland Regional Water Board, Auckland Regional Authority, by DSIR, Water Quality Centre, New Zealand.